gesis Leibniz Institute for the Social Sciences

Assessing Data Quality in the Age of Digital Social Research: A Systematic Review

Jessica Daikeler*, Leon Fröhling*, Lukas Birkenmaier*, Tobias Gummer*, Clemens Lechner*, Jan Schwalbach*, Indira Sen**, Henning Silber*, Bernd Weiss*, Katrin Weller*

18.03.2024, WEB DATA OPP, Barcelona, Spain

* GESIS – Leibniz Institute for the Social Sciences
** University of Konstanz, Germany

Leibniz Association

Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union NextGenerationEU

KODAQS

Relevance & Research Objectives

VIEWPOINT: THE FUTURE

Through the Glass Lightly

A collection of scientists at the frontier were asked what they see in the future for science.*
Here are their views....

> If you can look into the seeds of time,
> And say which grain will grow and which will not,
> Speak then to me, who neither beg nor fear
> Your favors nor your hate.
>
> Shakespeare, *Macbeth*, 1.3.58–61

THERE WILL BE ENORMOUS INROADS INTO human biology and human disease via genomics, gene therapy, and mouse knockout models; a revolution in drug design by

individuals at risk for diabetes, schizophrenia, obesity, and many other diseases. In many cases, dis-

toxins, sunlight, and so forth. The output will be a color movie in which the embryo develops into a fetus, is born, and then grows into an adult, explicitly depicting body size and shape and hair, skin, and eye color. Eventually the DNA sequence base will be expanded to cover genes important for traits such as speech and musical ability; the mother will be able to hear the embryo—as an adult—speak or sing.

Harvey F. Lodish
Whitehead Institute for Biomedical Research
Cambridge, Massachusetts

ILLUSTRATIONS BY TERRY E. SMITH

Weintraub, Hal. 1995. “Through the Glass Lightly.” *Science* 267(5204):1609–18. doi: 10.1126/science.7886446.

Designed vs. Found Data

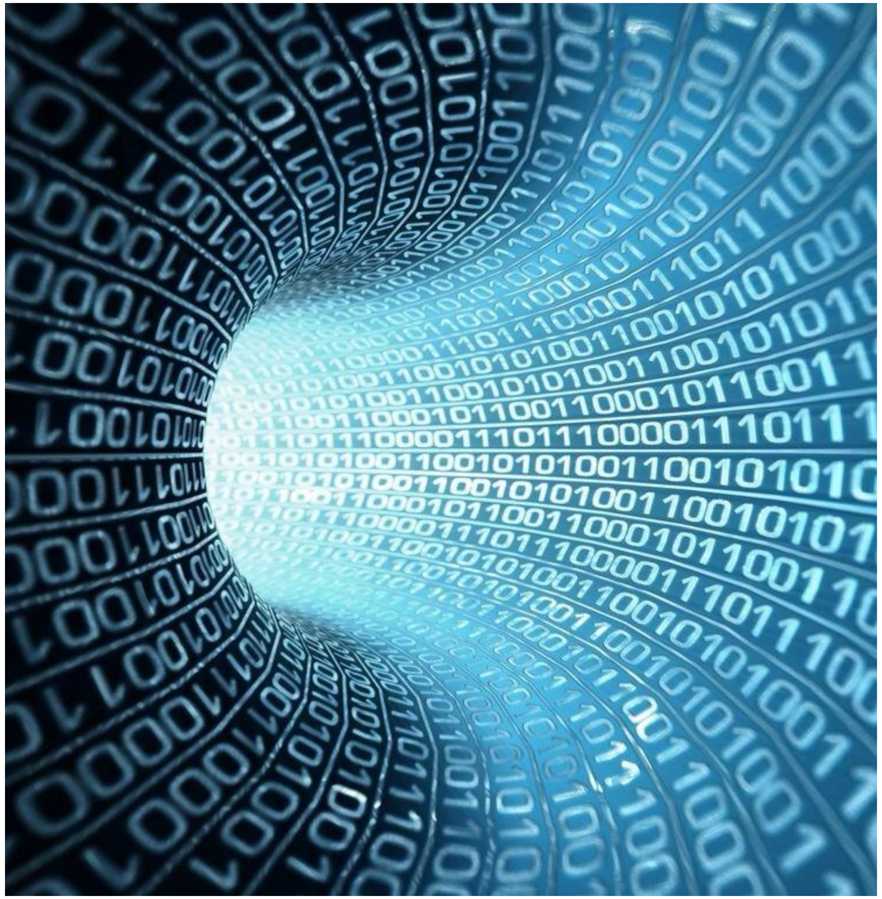
- **Designed data**: Data, e.g., survey questions designed with a pre-specified purpose in mind and to be representative for a specific target group. Since designed data are created with a pre-specified purpose the ratio of information to data is very high.
- **Found data/ Organic data**: Society has created systems that automatically track transactions of all sorts, data is created "organically" and has become an abundant, accessible and cheap commodity, e.g., tweets, images, videos, sensor data. Low information to data ratio.

Designed data	Organic data
Representative without information gaps but selective	Representative with information gaps but non-selective
Intrusive	Non-intrusive
Costly	Cheap
High information to data ratio	Low information to data ratio
Information on opinions, aspirations, preferences, actions planned and past actions	Information on transactions, actions, behavior, sentiment.

Source: https://norstatgroup.com/blog/why-do-we-need-surveys-when-we-have-access-to-so-much-data

What about data quality?

- "Data quality relates to the degree to which a set of inherent characteristics of data *(ISO 8000-2:2020)* fulfills intended operational decision-making and other specific roles *(Herzog, Scheuren, and Winkler 2007)."*
- Often systematized through so-called data quality or error frameworks
- Computational Social Science data quality = Social Science data quality concepts + Computer Science data quality concepts



Views on data quality

- *Extrinsic perspective:* Data is FAIR
  -> findable, accessible, interoperable, and reusable
- *Intrinsic perspective:* Data is accurate and complete to lead to the best possible evidence
- **Aim**: Systematize social science data quality concepts in the light of old and new social science research data

Our four objectives

I. We will provide researchers with a decision tree to identify the most appropriate data quality framework for a given use case.

II. We will determine which social science data types and quality dimensions are already addressed in the existing frameworks.

III. Considering different data types, we will identify gaps that are not yet covered by existing quality frameworks, and that should be addressed by future research.

IV. We will provide a detailed literature overview on data quality.

Data & Methods

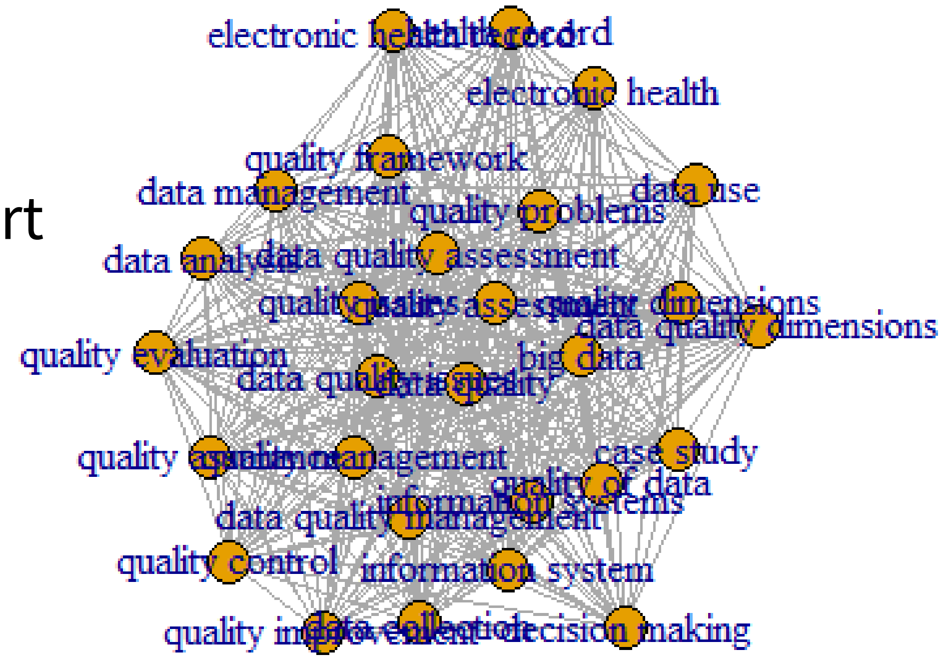
Methods

- Present our results with the help of a systematic review (objective 1, 2 & 4) and an evidence gap map (objective 1, 2 & 3).
- Rigorous methodological approach for systematic reviews *(Hedges and Cooper 2009)* and systematic approach described in *Grant and Boot (2009)* for evidence gap map

Source: https://egmopenaccess.3ieimpact.org/evidence-maps/gesis-survey-methods-evidence-map

Text Mining helped with Literature search

- litsearchr R Package *(GRames et al. 2019)* :
- Training search ("data quality" OR "error" OR "bias")
  AND ("framework" OR "concept" OR "perspective")
  in engines: Web of Science and Ebsco + Export
- Import training search result
- Extract keywords, titles and abstracts
- Get potential search terms
- Remove duplicates
- Group potential terms manually
- Search string will automatically be created

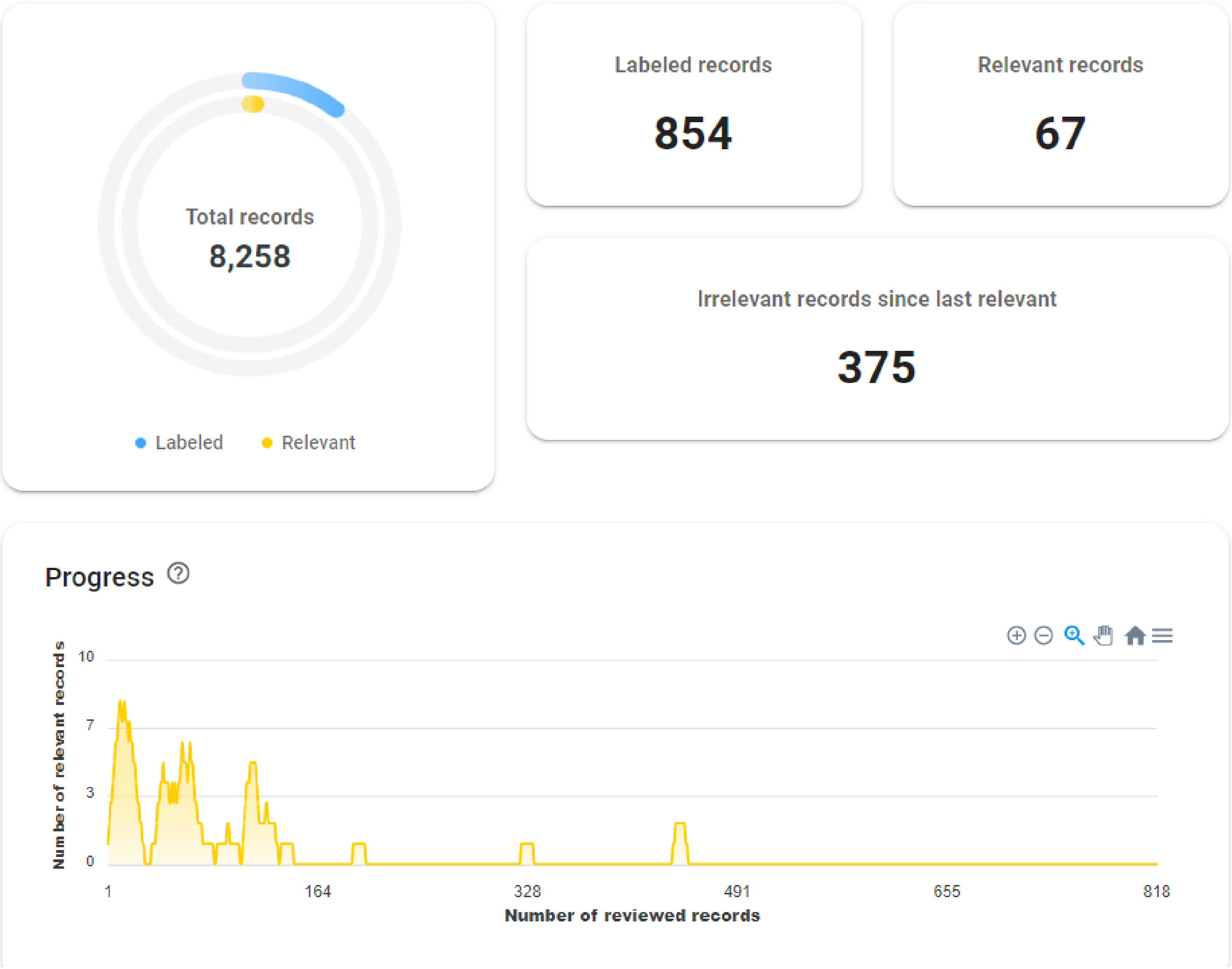


\(\(error* OR bias* OR "data* accuraci*" OR "data* analysi*" OR "data* clean*" OR "data* collect*" OR "data* complet*" OR "data* qualiti*" OR "data* valid*" OR "inform* qualiti*" OR "qualiti* assess*" OR "qualiti* assur*" OR "qualiti* improv*" OR "qualiti* of data*" OR "qualiti* evalu*"\) AND \(survey* OR "digit* content*" OR "digit* behavior*" OR poll* OR "public* opinion*" OR "big data*" OR "health* care*" OR "sensor* network*" OR "social* media*" OR "geograph* inform*" OR "wireless* sensor*"\) AND \(concept* OR "assess* framework*" OR "generic* framework*" OR "literatur* review*" OR "qualiti* dimens*" OR "qualiti* framework*" OR "qualiti* monitor*" OR "qualiti* problem*" OR "qualiti* requir*"\)\)

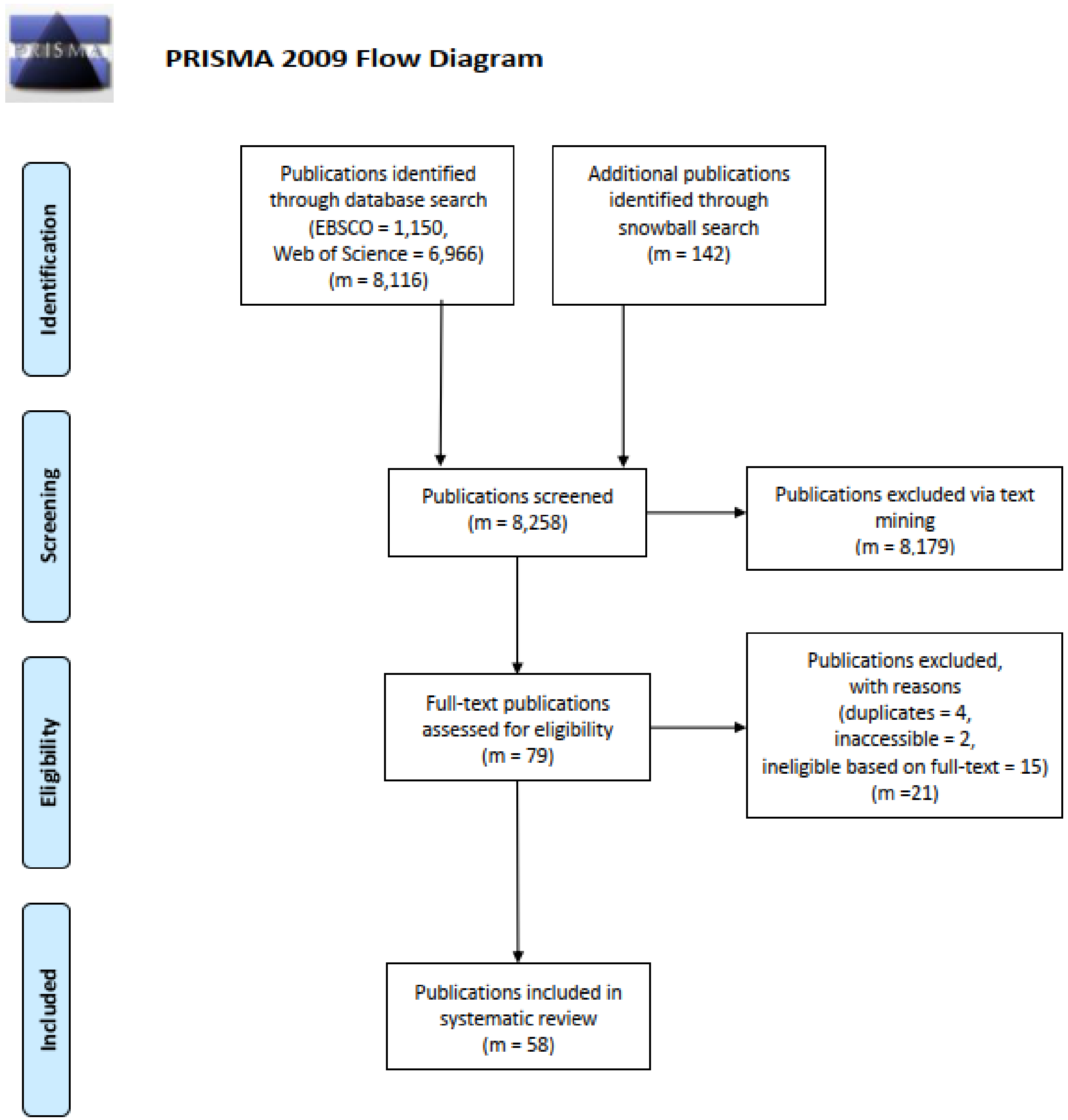
Literature screening

“ASReview” Python lab (https://asreview.nl/, *Van de Schoot et al. 2021*)

- Assists in screening literature
- Trains screening model based on example eligible and ineligible coding
- Displays the most likely eligible study next
- After deduplication N=58 eligible studies



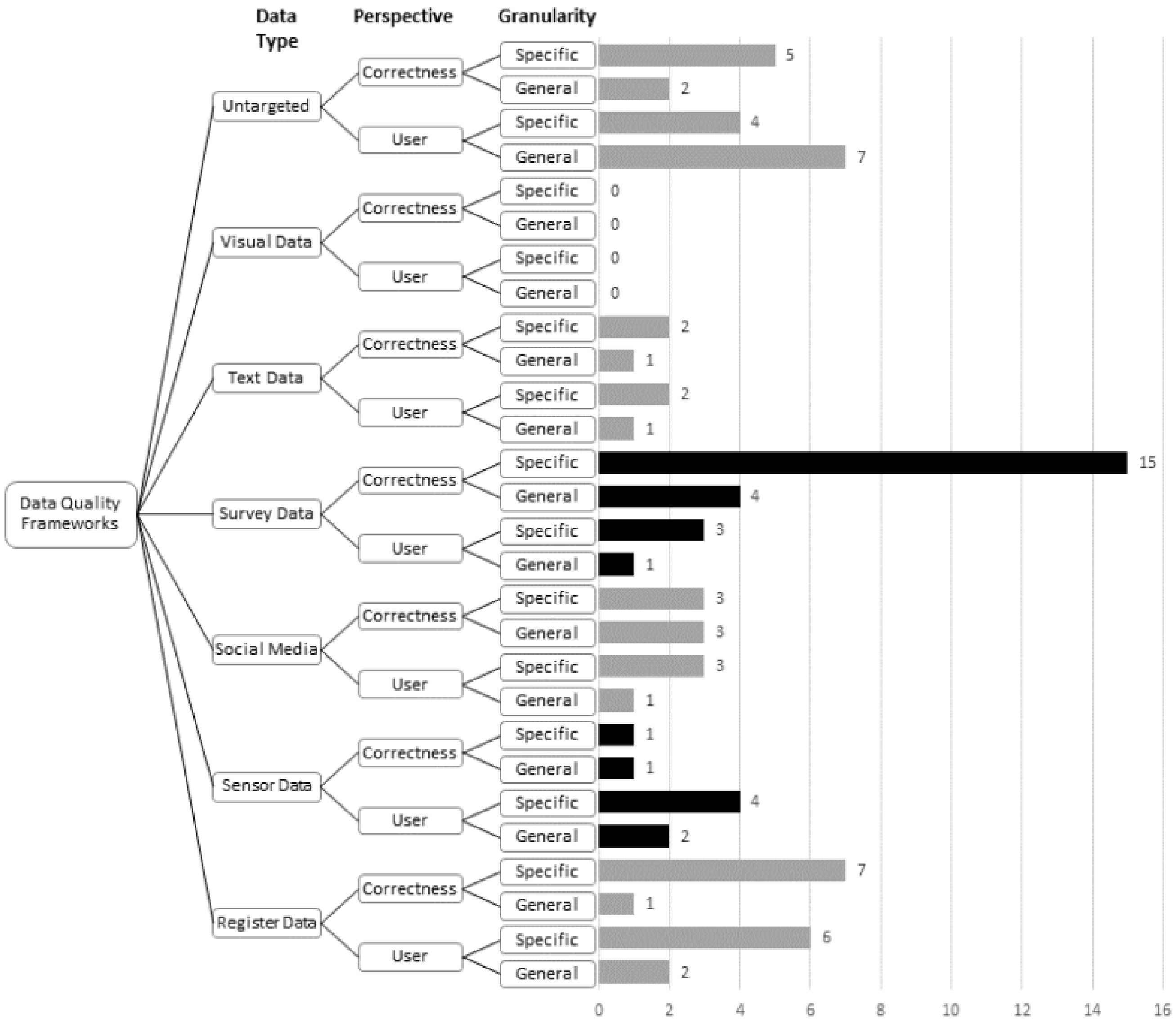
Final Study Sample



Results

Decision Tree

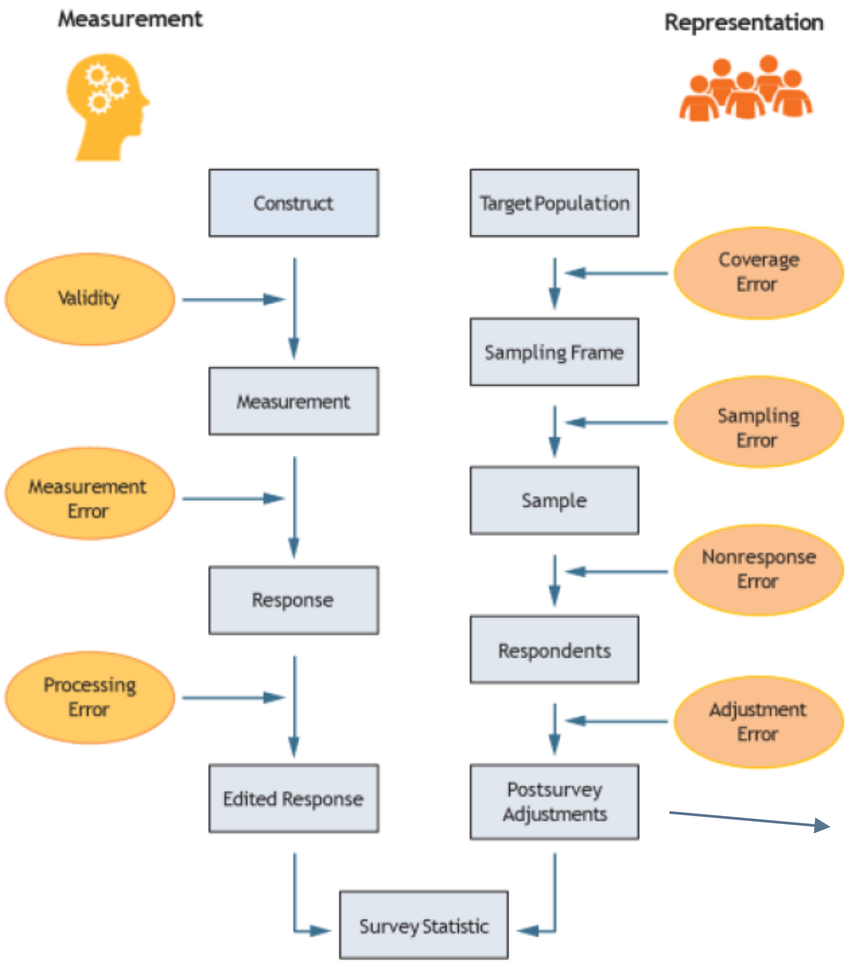
Soon Inter-active



TSE revisited?

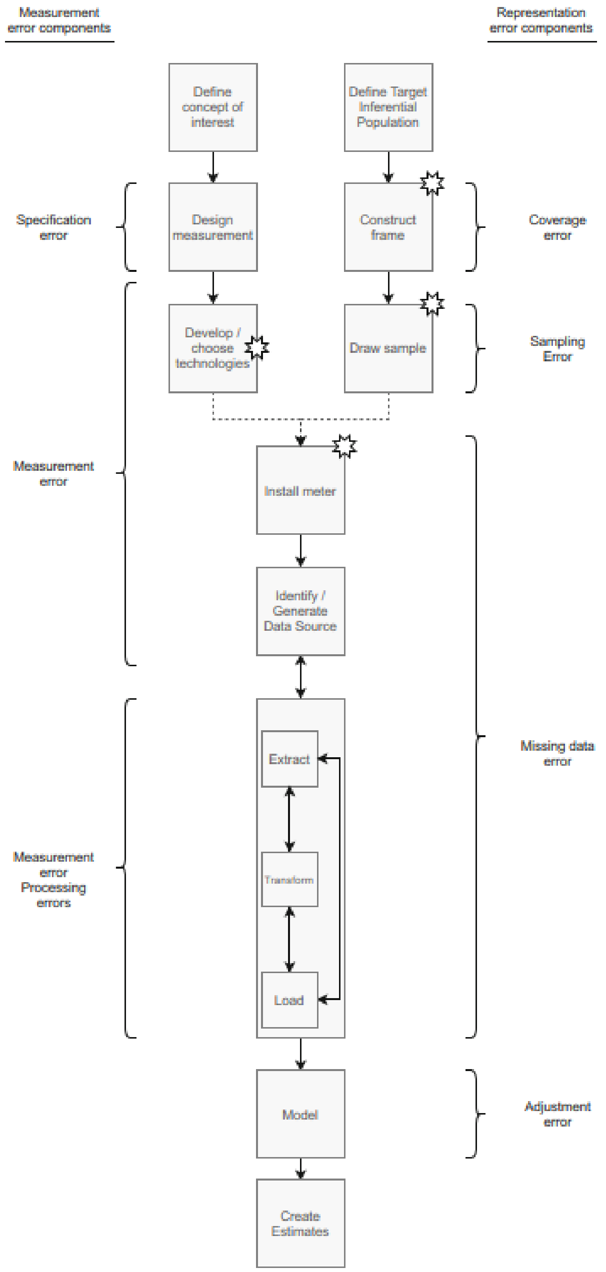
Measure-ment Error Response

Measure-ment Error Platform



(Groves and Lyberg 2010)

Model-ling Error

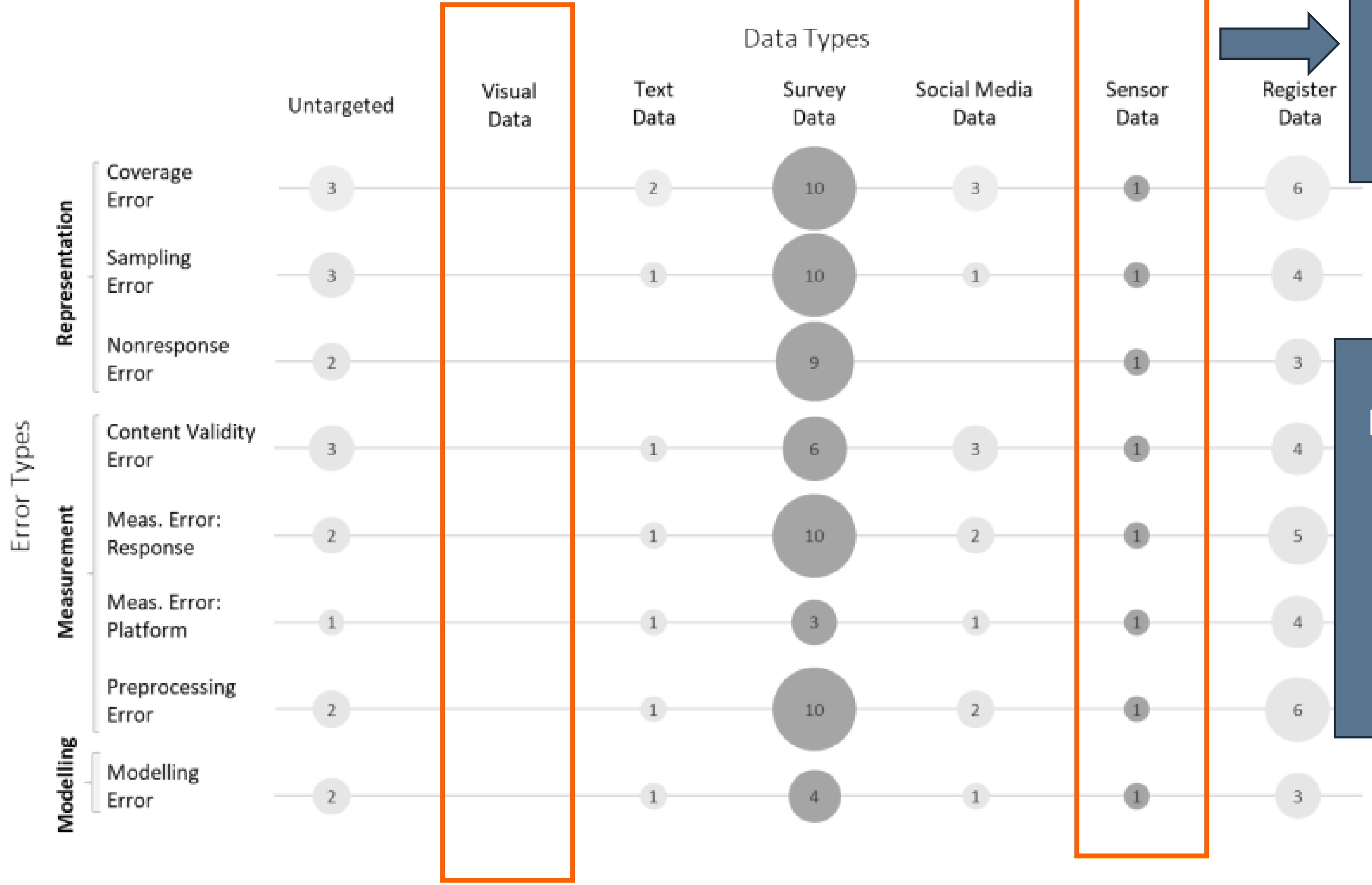


(Bosch and Revilla 2021)

Evidence Map

Inter-active

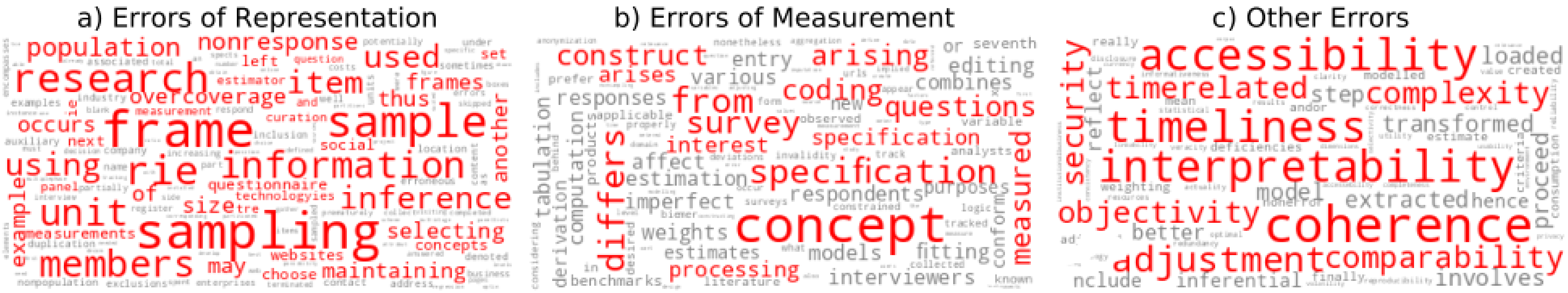
N= 39



Sensors = meters = browsing and app tracking behavior

Nothing found for other sensors e.g., gps, voice/ video recordings

Differences in Terminology!



Conclusion and Outlook

Conclusion

- **Two major perspectives** on data quality observable
  - Intrinsic: Error framework perspective
    - "Understand errors / biases in the data collection process"
  - Extrinsic: Usability / data characteristics perspective
    - "Evaluate the usability of data in relation to different quality characteristics (e.g., FAIR)"

- **Co-existing** of many frameworks: considerable variation in **data type(s)**; **dimensions** of data quality they cover and from which **perspective ->** systematic overview enables researchers to make informed fit-for-purpose decisions
- Different **disciplines**: Closer exchange of ideas between disciplines to ensure the proper implementation and advancement of research methods (e.g., difference terminology)
- **Research Gaps:**
  - **Linked data**: TSE likely approaches fall short in including all relevant data quality dimensions, but new approaches emerge (e.g., Christen, P., & Schnell, R. (2023). Thirty-three myths and misconceptions about population data: From data capture and processing to linkage. International Journal of Population Data Science, 8(1).)
  - **Addressing diverse sensor types**

Limitations and Outlook

- Frameworks stem mostly from Social and Computer Science (e.g., no biomarker medical literature, gps geography literature found)
- No evaluation of fit-for-purpose for existing frameworks
- Data quality indicators should be collected from the identified frameworks
  - Check KODAQS out : KODAQS
  - https://tinyurl.com/kodaqsdataquality

gesis Leibniz Institute for the Social Sciences

References

Bosch, Oriol J., and Melanie Revilla. 2021. “When Survey Science Met Online Tracking : Presenting an Error Framework for Metered Data.”

Christen, P., & Schnell, R. (2023). Thirty-three myths and misconceptions about population data: From data capture and processing to linkage. International Journal of Population Data Science, 8(1).

Grames, Eliza M., et al. "An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks." Methods in Ecology and Evolution 10.10 (2019): 1645-1654.

Grant, Maria J., and Andrew Booth. 2009. “A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies.” Health Information & Libraries Journal 26(2):91–108. doi: 10.1111/j.1471-1842.2009.00848.x.

Groves, Robert M., and Lars Lyberg. 2010. “Total Survey Error: Past, Present, and Future.” Public Opinion Quarterly 74(5):849–79.

Hedges, LV, and H. Cooper. 2009. “Research Synthesis as a Scientific Process.” The Handbook of Research Synthesis and Meta-Analysis 1:4–7.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler. 2007. “What Is Data Quality and Why Should We Care?” Pp. 7–15 in Data quality and record linkage techniques. Springer.

ISO 8000-2:2020. n.d. Data Quality. International Organization for Standardization, Geneva, Switzerland.

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. “A Total Error Framework for Digital Traces of Human Behavior on Online Platforms.” Public Opinion Quarterly 85(S1):399–422. doi: 10.1093/poq/nfab018.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., ... & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, 3(2), 125-133.

Weintraub, Hal. 1995. “Through the Glass Lightly.” *Science* 267(5204):1609–18. doi: 10.1126/science.7886446.

Citation

Daikeler, Jessica, Leon Fröhling, Indira Sen, Lukas Birkenmaier, Tobias Gummer, Jan Schwalbach, Henning Silber, Bernd Weiß, Katrin Weller, and Clemens Lechner. 2024 (Forthcoming). "Assessing Data Quality in the Age of Digital Social Research: A Systematic Review." *Social Science Computer Review*.

Thank you for your attention

Leibniz Institute
for the Social Sciences

jessica.daikeler@gesis.org

JeSs_Dalk & _kodaqs_

Appendix

Eligibility criteria

1. **Data Quality and Error:** The contribution needs to explicitly address data quality or error concepts (or synonymous).
2. **Concept:** The contribution needs to characterize their work as a concept or synonymous (no primary studies).
3. **Social Science data:** The contribution needs to explicitly elaborate on Social Science data.
4. **Human Beings:** The framework should have a focus on the observation of human beings.
5. **Data type:** The contribution should target on survey and online content data (e.g., text, images, videos) as those two are widely used.
6. **Data collection:** Data collections in digital and offline scenarios are eligible.
7. **Researcher perspective:** Contributions visiting data quality from an archive / data management perspective by elaborating on archiving strategies (e.g., FAIR) are not eligible for our study.
8. **Published**: Contribution needs to be published (no grey literature).

Coding scheme example

ID	Authors	Title	Year	DOI	TARGET	Survey I	Registe	Text Da	Video D	Images	Sensor I	Web Tra	Social M	DATATYPE	Perspec	Else	PERSPECTIVE	Da
SB1127	Groves, Robert	Survey M	2009	https://ebc	1	1	0	0	0	0	0	0	0	Survey Data	1	99	Data Perspective	
SB1128	Kish, Leslie	Survey S	1965	https://psy	1	1	0	0	0	0	0	0	0	Survey Data	3	99	Data Perspective	
DB1779	Micic, N; Neagu	Towards	2017	https://doi	1	0	0	0	0	0	1	0	0	Sensor Data	4	99	User Perspective	Se
DB8246	Bodell, Miriam	From Do	2022	https://doi	1	0	0	1	0	0	0	0	0	Text Data	2//4	99	Perspective else: Data and User Perspective	n.a
DB7661	Merino, J; Caba	A Data C	2016	https://doi	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Co
DB8201	Japec, Lilli; Kreu	Big Data	2015	https://doi	1	1	0	0	0	0	0	0	0	Survey Data	1	99	Data Perspective	
DB8253	Agarwal, Nitin;	Informat	2010	https://ww	1	0	0	0	0	0	0	0	1	Social Media Data	4	99	User Perspective	Ac
DB8216	Olteanu, Alexai	Social Da	2019	https://doi	0	0	0	0	0	0	0	0	0	untargeted	2	99	Data Perspective	
SB1005	Guptill, S.C.; Mc	Element	1995	https://ww	1	0	1	0	0	0	1	0	0	Register DataSensor Data	4	99	User Perspective	Ac
SB1082	Radhakrishna, F	Ensuring	2012	https://eri	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Va
SB1086	Baur, Nina	Measure	2009	https://nbr	1	0	1	0	0	0	0	0	0	Register Data	3	99	Data Perspective	
SB1101	Deming, E.	On Error	1944	https://doi	1	1	0	0	0	0	0	0	0	Survey Data	1	99	Data Perspective	
SB1120	Eurostat	Handboo	2007	https://par	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Re
SB1123	European Statis	Quality /	2012	https://ec.	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Re
SB1126	Daas, P.J.H.; Are	Quality I	2008	http://ww	1	0	1	0	0	0	0	0	0	Register Data	2	99	Data Perspective	
SB1136	Brown, Paul A; /	A methc	2022	https://doi	0	0	0	0	0	0	0	0	0	untargeted	2//4	99	Perspective else: Data and User Perspective	n.a
DB8228	Hsieh, YP; Murp	Total Tw	2017	http://doi.	1	0	0	0	0	0	0	0	1	Social Media Data	1	99	Data Perspective	
DB8153	Tufekci, Z.	Big Ques	2014	https://ww	1	0	0	0	0	0	0	0	1	Social Media Data	1	99	Perspective else: Data and User Perspective	
DB7159	Lynn, T; Kilroy, !	Towards	2015	https://iee	1	0	0	0	0	0	0	0	1	Social Media Data	1	99	Perspective else: Data and User Perspective	
DB3706	Holtom, B; Baru	Survey r	2022	https://doi	1	1	0	0	0	0	0	0	0	Survey Data	1	99	Data Perspective	
DB1331	Hong, JH; Huang	Enabling	2017	https://doi	1	0	1	0	0	0	1	0	0	Register DataSensor Data	4	99	User Perspective	Co
DB7326	Juddoo, S	Overview	2015	https://iee	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Co
DB7337	Lukyanenko, R;	Expectin	2019	https://doi	1	0	0	1	0	0	0	0	0	Text Data	4	99	User Perspective	Ac
DB7393	Ijab, MT; Surin,	Concept	2019	https://doi	0	0	0	0	0	0	0	0	0	untargeted	4	99	User Perspective	Av
DB7906	Kimberlin, CL; V	Validity	2008	https://doi	1	1	1	0	0	0	1	0	0	Survey DataRegister DataSensor Da	1	99	Data Perspective	
SB1002	Devillers, R; Bé	Multidin	2005	https://doi	1	0	1	0	0	0	1	0	0	Register DataSensor Data	4	99	Perspective else: Data and User Perspective	
SB1011	Yang, T.	Visualisa	2007	https://doi	1	0	1	0	0	0	1	0	0	Register DataSensor Data	4	99	Perspective else: Data and User Perspective	
SB1078	Batini C; Rula A;	From Da	2015	https://doi	1	0	1	1	0	0	1	0	0	Register DataText DataSensor Data	4	99	User Perspective	Ac

What we found!